

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号
特開2002-108887
(P2002-108887A)

(43)公開日 平成14年4月12日 (2002.4.12)

(51)Int.Cl. ⁷ G 0 6 F 17/30	識別記号 2 1 0 1 7 0 3 2 0	F I G 0 6 F 17/30	テ-マート(参考) 2 1 0 A 5 B 0 7 5 1 7 0 A 3 2 0 C
-------------------------------------------	---------------------------------	----------------------	------------------------------------------------------

審査請求 未請求 請求項の数11 O.L (全 8 頁)

(21)出願番号 特願2000-299971(P2000-299971)

(22)出願日 平成12年9月29日 (2000.9.29)

(71)出願人 000001007
キヤノン株式会社
東京都大田区下丸子3丁目30番2号

(72)発明者 中越 治樹
東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

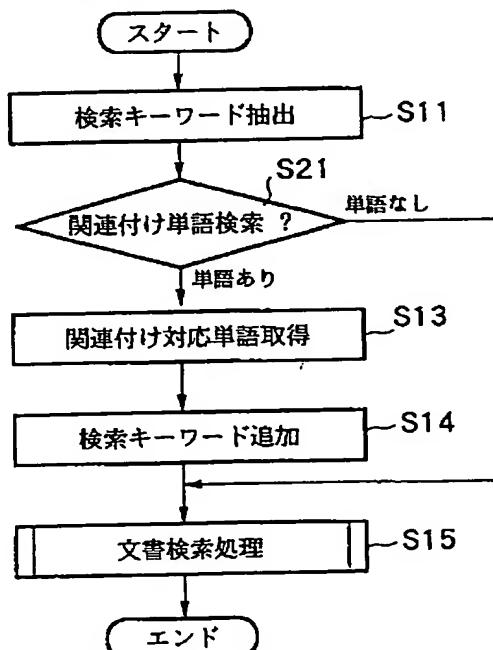
(74)代理人 100076428
弁理士 大塚 康徳 (外2名)
Fターム(参考) 5B075 ND03 NK02 NK31 PP22

(54)【発明の名称】 文書検索装置、その装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体

(57)【要約】

【課題】 未知のキーワードに対しても効率的な文書検索を行う。

【解決手段】 設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索するに際して、設定されたキーワードに基づいて、関連付け辞書を参照することにより、当該キーワードに関連付けされている単語を他のキーワードとして追加設定し、それら複数のキーワードに従って、検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する(S11-S15)。ここで、関連付け辞書は、入力文書から抽出した単語が、予め記憶されているキーワード辞書に登録されていない未知語である場合に、その入力文書における当該未知語の前方または後方に位置する単語とを関連付けした辞書であり、S11-S15に先立って作成される。



【特許請求の範囲】

【請求項1】 設定されたキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置であって、

入力文書から単語を抽出し、抽出した単語が第1の辞書に登録されていない未知語か否かを判断する未知語検出手段と、

前記未知語検出手段によって未知語であると判断された単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け手段と、

設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加手段と、を備えることを特徴とする文書検索装置。

【請求項2】 更に、前記キーワード追加手段によって展開された複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する検索手段を備えることを特徴とする請求項1記載の文書検索装置。

【請求項3】 前記検索手段は、前記未知語検出手段によって未知語が検出されないときに、前記検索対象の文書の中から、前記設定されたキーワードを含む文字列だけを検索することを特徴とする請求項2記載の文書検索装置。

【請求項4】 前記未知語検出手段及び前記関連付け手段は、前記キーワード追加手段及び前記検索手段の実行タイミングとは別個に、前記検索対象の文書に対して定期または不定期に実行されることを特徴とする請求項1または請求項2記載の文書検索装置。

【請求項5】 前記キーワード追加手段は、前記設定されたキーワード、前記第2の辞書を参照することによって追加されたキーワード、並びに前記第1の辞書に基づいて、それらキーワードを更に追加設定する手段を含むことを特徴とする請求項1記載の文書検索装置。

【請求項6】 設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置へのキーワード追加方法であって、

入力文書から単語を抽出し、抽出した単語が第1の辞書に登録されていない未知語か否かを判断する未知語検出手段と、

前記未知語検出手段にて未知語であると判断した単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け工程と、

設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加工程と、を有することを特徴とするキーワード追加方法。

10

20

30

40

50

【請求項7】 請求項6記載のキーワード追加方法によって展開した複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索することを特徴とする文書検索方法。

【請求項8】 前記検索工程では、前記未知語検出手段にて未知語を検出しなかったときに、前記検索対象の文書の中から、前記設定されたキーワードを含む文字列だけを検索することを特徴とする請求項7記載の文書検索方法。

【請求項9】 請求項1乃至請求項5の何れかに記載の文書検索装置としてコンピュータを動作させるプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項10】 請求項6記載のキーワード追加方法をコンピュータによって実現可能なプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項11】 請求項7または請求項8記載の文書検索方法をコンピュータによって実現可能なプログラムコードが格納されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、検索対象として入力されたキーワードに基づいて、そのキーワードを含むまたは関係する文書を効率良く出力する文書検索装置に関する。

【0002】

【従来の技術】 従来より、コンピュータを用いて、オペレータによって入力されたキーワードを含む文書を、予め記憶されている複数の文書の中から検索する技術が提案されている。

【0003】 また、近年においては、あるキーワードに対して関連付けされた他の単語が予め登録されている辞書（以下、キーワード辞書）を用意しておき、オペレータによって入力されたキーワードに関連性のある単語を当該キーワード辞書から取得し、文書の検索に際しては、入力されたキーワードに加え、当該キーワード辞書から取得した単語を他のキーワードとして併せて利用する技術も提案されている。

【0004】

【発明が解決しようとする課題】 上記のキーワード辞書を利用する文書検索処理によれば、そのキーワード辞書にオペレータによって入力されたキーワードが含まれる場合、検索対象の文書に当該キーワードが含まれていなくても、実際には関連性がある文書をもヒットすることができ、文書検索を効率的に行うことができる。

【0005】 しかしながら、入力されたキーワードがキーワード辞書に登録されていない未知の単語（以下、未

知語) である場合には、そのキーワードを直接的に含む文書しかヒットすることはできない。このため、キーワードとして新語や未知語が入力された場合には、効率的な文書検索を行うことができない。

【0006】そこで本発明は、未知のキーワードに対しても効率的な文書検索を行う文書検索装置及びその装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体の提供を目的とする。

【0007】

【課題を解決するための手段】上記の目的を達成するため、本発明に係る文書検索装置は、以下の構成を特徴とする。

【0008】即ち、設定されたキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索装置であって、入力文書から単語を抽出し、抽出した単語が第1の辞書(キーワード辞書)に登録されていない未知語か否かを判断する未知語検出手段と、前記未知語検出手段によって未知語であると判断された単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書を作成する関連付け手段と、設定されたキーワードに基づいて前記第2の辞書(関連付け辞書)を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加手段と、前記キーワード追加手段によって展開された複数のキーワードに従って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索する検索手段とを備えることを特徴とする。

【0009】好適な実施形態において、前記未知語検出手段及び前記関連付け手段は、前記キーワード追加手段及び前記検索手段の実行タイミングとは別個に、前記検索対象の文書に対して定期または不定期に実行される。

【0010】また、上記の同目的を達成するため、本発明に係る文書検索装置へのキーワード追加方法は、以下の構成を特徴とする。

【0011】即ち、設定したキーワードに基づいて、検索対象の文書の中から、そのキーワードを含む文字列を検索する文書検索方法であって、入力文書から単語を抽出し、抽出した単語が第1の辞書(キーワード辞書)に登録されていない未知語か否かを判断する未知語検出工程と、前記未知語検出工程にて未知語であると判断した単語と、前記入力文書における該単語の前方または後方に位置する単語とが関連付けされた第2の辞書(関連付け辞書)を作成する関連付け工程と、設定されたキーワードに基づいて前記第2の辞書を参照することにより、該キーワードに関連付けされている単語を新たなキーワードとして追加設定するキーワード追加工程を有することを特徴とする。

【0012】また、文書検索方法であって、上記のキーワード追加方法によって展開した複数のキーワードに従

って、前記検索対象の文書の中から、それらキーワードのうち少なくとも1つを含む文字列を検索することを特徴とする。

【0013】更に、上記の文書検索装置、その装置へのキーワード追加方法、並びに文書検索方法を、コンピュータによって実現するプログラムコードが格納されている、コンピュータ読み取り可能な記憶媒体を特徴とする。

【0014】

【発明の実施の形態】以下、本発明の一実施形態を図面を参照して説明する。はじめに、本発明に係る文書検索装置を実現可能なコンピュータの構成を図13を参照して説明する。

【0015】図13は、本発明を適用可能なコンピュータのプロック構成図である。

【0016】図中、22は、CRT等のディスプレイ、23は検索用のキーワードや文書等の入力手段であるキーボードである。24は、ブートプログラム等を記憶しているROMである。25は、各種処理結果を一時記憶するRAMである。26は、検索対象となる複数の文書(文書データ)からなるデータベースや、後述する文書検索処理を実現するプログラム等を記憶するハードディスクドライブ(HDD)等の記憶装置である。27は、外部装置から文書データ等を受信可能な通信インターフェースである。そして28は、処理結果等を印刷するプリンタである。これらの各構成は、内部バス29を介して接続されており、CPU(中央演算処理装置)21は記憶装置26に記憶したプログラムに従って、文書検索装置としてのコンピュータの全体を制御する。

【0017】尚、複数の文書(文書データ)からなるデータベース(D/B)は、外部装置に記憶されており、その外部装置に通信インターフェース27を介してアクセスする構成であっても良い。

【0018】次に、図13に示すコンピュータによって実現される文書検索の処理手順について説明する。本実施形態に係る文書検索処理は、大別して、以下に説明する単語インデックス作成処理(図1)と、キーワード追加(展開)工程を含む文書検索処理とからなる。

【0019】ここで、単語インデックス作成処理は、後述する関連付け辞書を更新すべく、例えばオペレータの指示に応じて、記憶装置26に記憶されたD/B内の文書群を対象として、定期または不定期に実行されるメンテナンスとしての位置付けである。一方、文書検索処理は、オペレータによってキーワードが入力されたときに、当該単語インデックス作成処理によって既に作成

(更新)されている関連付け辞書を参照しながら、入力されたキーワードを含む及び/または関連性のある文書(文字列)を記憶装置26に記憶されたD/B内の文書群の中から実際に抽出する処理である。

【0020】尚、本実施形態では、説明を容易にすべ

く、上記の単語インデックス作成処理(図1)において、関連付け辞書に登録されていない単語として「トルシオ」を例に挙げ、以下に説明する文書検索処理(図2)においても、入力されるキーワードを「トルシオ」としたため、単語インデックス作成処理において注目した図3の文書1乃至3における注目箇所(実際に文字列が示された部分)と、文書検索処理の検索結果を示す図12の文書1乃至3における注目箇所と同じであるが、説明の都合からあり、実際には、入力されるキーワードに応じて、検索対象のD/B内から好みの文書が選択される。

【0021】<単語インデックス作成処理>図1は、本実施形態における単語インデックス作成処理を示すフローチャートである。

【0022】同図において、ステップS1の単語抽出処理では、図3に示すような文書群(文書1乃至3)を抽出対象として、それら文書の文字列から、その文字列を構成する単語を抽出する。本ステップにおいて単語を抽出する際には、予め記憶されている図6に示すようなキーワード辞書を参照することにより、そのキーワード辞書に既に登録されている単語が、単語抽出対象の各文字列に含まれているか否かを検索する。

【0023】このとき、単語抽出対象の文字列の中に、当該キーワード辞書に登録されていない単語が含まれていた場合には、その単語の前方または後方に連続する文字種などを判断して当該単語を抽出する。より具体的には、図3に示す文書1を例にすると、キーワード辞書を参照することにより、図6には不図示の「監督」、「指導方法」、「頭脳的」などの単語が抽出され、当該キーワード辞書に登録されていない単語「トルシオ」は、カタカナが連続する単語として判断されて、文書1の文字列の中から抽出される。

【0024】ステップS2の未知語検出処理では、ステップS1で抽出した各単語の中に、キーワード辞書に登録されていない未知語が存在するかどうかを判断し、この判断でYES(未知語有り)のとき(上記の例の場合は、単語「トルシオ」が未知語)にはステップS3に進み、NO(未知語無し)のときにはステップS5に進む。

【0025】ステップS3の関連付け作成処理では、ステップS2で検出された未知語と、その未知語が抽出された文字列における当該未知語の前方または後方に存在する単語とを関連付けする。上記の例の場合、図3に示す文書1において、未知語として検出された「トルシオ」に、その単語に後続する「監督」が関連付けされる。本ステップにて関連付けが作成できたときにはステップS4に進み、作成できないときにはステップS5に進む。

【0026】ステップS4の関連付け保存処理では、ステップS3で作成した関連付けを、関連付け辞書とし

て、RAM25や記憶装置26に保存する。このような関連付けが単語抽出対象の各文字列を対象にして行われることにより、図3に示す3つの文書の場合には、関連付け辞書に、図9に示すように未知語「トルシオ」に「監督」と「ジャパン」とが関連付けされた状態で新たに保存される。

【0027】ステップS5の単語インデックス作成処理では、ステップS1で抽出された各単語(未知語を含む)を文書毎にまとめるこにより、単語インデックスが作成される。上記の例の場合には、図3に示す文書1乃至3から抽出された各単語により、図5に示す3つの単語インデックスが作成される。これらの単語インデックスは、その単語インデックスの作成の元となった文書データとD/B内で関連付けされた状態で保存され、後述する文書検索処理(図2)のステップS15において参照される。

【0028】<文書検索処理>図2は、本実施形態における文書検索処理を示すフローチャートである。

【0029】同図において、ステップS11の検索キーワード取得処理では、オペレータによってキーボード23等から入力されるキーワードを取得する。ここでは、図4に示すように、キーワードとして「トルシオ」が取得されるものとする。

【0030】ステップS12の関連付け単語検索処理では、ステップS11で入力されたキーワードが、関連付け辞書に現時点で既に登録されているか否かを検索し、この検索の結果、既に登録されているときにはステップS13に進み、未登録のときには、検索対象の文書の中にステップS11にて取得したキーワードを直接含む文書を検索すべくステップS15に進む。上記の例の場合は、図1を参照して上述した単語インデックス作成処理により、図9に示す関連付け辞書には、キーワード「トルシオ」が既に登録されている。

【0031】ステップS13の関連付け対応単語の取得処理では、ステップS12で検索された単語に関連付けされている単語(以下、対応単語)を取得する。上記の場合は、ステップS12において、キーワード「トルシオ」を、関連付け辞書の中からヒットすることができるので、当該関連付け辞書から「トルシオ」に関連付けされている「監督」及び「ジャパン」の2つの単語が対応単語として取得される。

【0032】ステップS14の検索キーワード追加処理では、ステップS13で取得した対応単語を、ステップS11にて入力されたキーワードに関連性を有する他のキーワードとして追加する。上記の場合は、ステップS11にてオペレータによって設定された図4のキーワード「トルシオ」に加えて、図10に示すように、「監督」及び「ジャパン」の2語が新たに検索用のキーワードとして追加されている。即ち、本実施形態では、元の検索キーワードであった「トルシオ」に加え、そのキーワードとして追加されている。

ワードに関連性のある複数のキーワードに展開されたことになる。

【0033】ステップS15：本ステップでは、上記のステップまでに用意された少なくとも1つのキーワードに従って、図1のステップS5にて記憶した単語インデックスを参照することにより、D/B内で当該単語インデックスに関連付けされている文書群を検索する。

【0034】即ち、ステップS11にて入力されたキーワードが関連付け辞書に登録されてなかった場合には、そのキーワードを直接含む文書（単語インデックス）を検索する一般的な検索処理を行う。この場合、入力されたキーワードが「トルシオ」であれば、検索結果として、図12に示す文書1及び2が得られる。

【0035】一方、ステップS14までの各ステップにより、ステップS11にて入力されたキーワードに加え、他のキーワードが追加された場合には、それらのキーワードが「トルシオ」、「監督」及び「ジャパン」であれば、それらのキーワードが含まれる単語インデックスにより、検索結果として、図12に示す文書1乃至3が得られる。

【0036】このように、本実施形態によれば、オペレータによって入力された検索キーワードを、図1の単語インデックス処理によって作成された関連付け辞書を利用して展開することにより、効率的な文書検索処理を行うことができ、オペレータによって入力されたキーワードを直接的には含まないものの、実施には関連性がある文書をもヒットすることができ、文書検索を効率的に行うことができる。

【0037】尚、ステップS15では、D/Bを検索するのに先立って、予め記憶されているキーワード辞書（例えば図6）と不図示の辞書とを参照することにより、キーワード展開辞書（図7の右側の例参照）を作成し、そのキーワード展開辞書を参照しながら、従来より知られているキーワード展開処理を行うことにより、上述したステップS14にて展開済みのキーワード（図11）を、図11に示すように、更に多くのキーワードに展開することができる。従って、これら複数のキーワードを用いて文書検索を行えば、入力されたキーワードを直接的には含まないものの、実際には関連性がある文書を、より高い確立でヒットすることができ、操作性が向上する。

【0038】

【他の実施形態】尚、本発明は、複数の機器（例えばホストコンピュータ、インターフェイス機器、リーダ、プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機、ファクシミリ装置など）に適用してもよい。

【0039】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記

録した記憶媒体（または記録媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出し実行することによって、達成される。この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているオペレーティングシステム（OS）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれる。

【0040】さらに、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張カードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張カードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれる。

【0041】

【発明の効果】以上説明した本発明によれば、未知のキーワードに対しても効率的な文書検索を行う文書検索装置及びその装置へのキーワード追加方法、文書検索方法及びコンピュータ読み取り可能な記憶媒体の提供が実現する。

【図面の簡単な説明】

【図1】本実施形態における単語インデックス作成処理を示すフローチャートである。

【図2】本実施形態における文書検索処理を示すフローチャートである。

【図3】関連付け辞書の作成のために参照する文書群を例示する図である。

【図4】関連付け辞書を例示する図である。

【図5】単語インデックスの作成例を示す図である。

【図6】キーワード辞書の例を示す図である。

【図7】キーワード展開辞書を例示する図である。

【図8】検索結果を例示する図である。

【図9】関連付け辞書の作成例を示す図である。

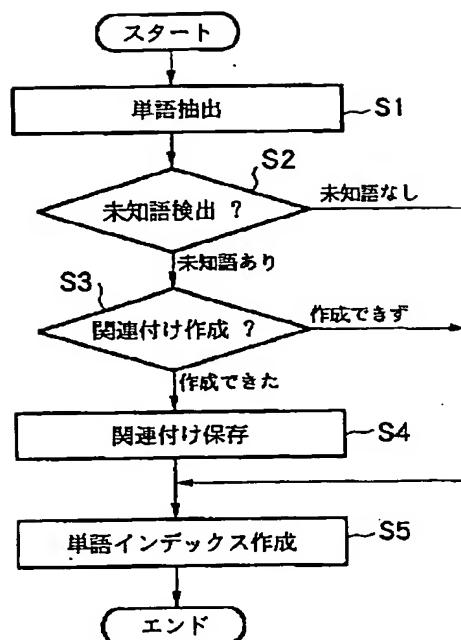
【図10】キーワードの展開例を示す図である。

【図11】図10に示すキーワードの展開例を、キーワード展開辞書（図7）を参照して更に展開した場合を例示する図である。

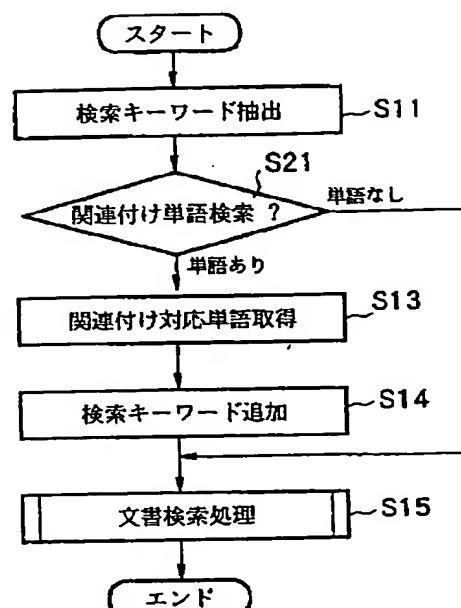
【図12】検索結果である文書群を例示する図である。

【図13】本発明を実現可能なコンピュータのブロック構成図である。

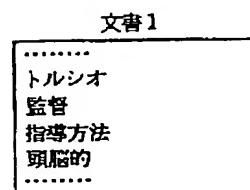
【図1】



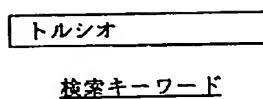
【図2】



【図5】

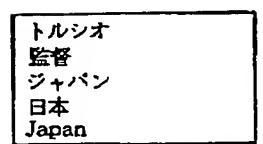


【図4】



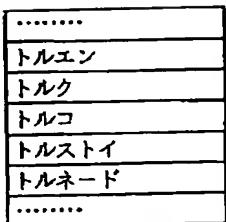
検索キーワード

【図11】



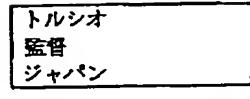
検索キーワード展開

【図6】



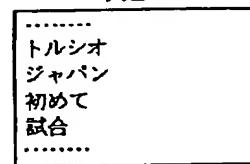
キーワード辞書

【図10】

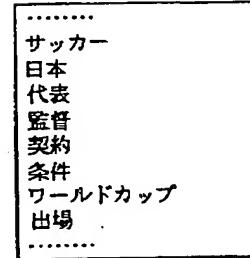


検索キーワード展開

文書2



文書3



単語インデックス群

【図3】

文書1
.....
トルシオ監督の指導方法は、頭脳的であり、
.....

文書2
.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

文書3

.....
サッカー日本代表監督の契約条件としては、
ワールドカップ出場であるが、
.....

【図7】

.....
ジャパン	日本 Japan
.....
トルエン	可燃性液体
トルコ	トルコ石 トルコ絨毯
.....

キーワード展開辞書

【図9】

.....
トルシオ	監督 ジャパン
トンガ	王国
.....

関連付け辞書

【図12】

文書1

.....
トルシオ監督の指導方法は、頭脳的であり、
.....

文書2

.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

【図8】

文書1

.....
トルシオ監督の指導方法は、頭脳的であり、
.....

.....
サッカー日本代表監督の契約条件としては、
ワールドカップ出場であるが、
.....

文書3

検索結果文書群

文書2

.....
トルシオジャパンとして、初めての試合が行
われたが、
.....

検索結果文書群

【図13】

